DOI: 10.1093/nar/gkh459

BLAST Filter and GraphAlign: rule-based formation and analysis of sets of related DNA and protein sequences

John B. Spalding[1],* and Peter J. Lammers[1,2]

[1]Molecular Biology Program and [2]Department of Chemistry and Biochemistry, New Mexico State University,
Las Cruces, NM 88003, USA

Received February 13, 2004; Revised and Accepted April 26, 2004

ABSTRACT

BLAST Filter and GraphAlign are web-based tools that offer novel methods for building and analyzing sets of related (i.e. similar) DNA and protein sequences. They can be used separately or together. BLAST Filter generates related sequence sets in an automated, objective and reproducible way based on an input query sequence. Sequences matched by BLAST are filtered through a set of 15 user-configurable rules based on full-length query/subject comparisons, high-scoring segment pair statistics and the level of redundancy in the sequence set. Such sets can be used for multiple alignments, profile hidden Markov models and other bioinformatics applications, including GraphAlign, which provides several novel methods for analyzing global query/subject alignments along with graphical representations of sequence similarities. These services are available at the following URLs: http://darwin.nmsu.edu/cgi-bin/blast_filter.cgi and http://darwin.nmsu.edu/cgi-bin/graph_align.cgi.

AVAILABILITY AND ADDITIONAL INFORMATION

BLAST Filter and GraphAlign are available at the URLs above and run in a distributed fashion on a 32-node Beowulf cluster at the Southwest Biotechnology and Informatics Center. Results are returned directly to the user's browser and may be refined iteratively through changes in rule and parameter values. Additional information is provided in extensive online Help pages.

INTRODUCTION

A common requirement in bioinformatics analyses is the need to build a set of similar DNA or protein sequences. Such sets are formed, e.g. as the first step in generating multiple alignments, molecular phylogenetic analyses (1), position-specific scoring matrices, profile hidden Markov models (2) and multiple sequence methods for protein secondary structure prediction (3). This presents two problems for the researcher: to create a consistent, well-defined method for selecting the sequences, and to obtain the sequences themselves. BLAST Filter addresses both of these problems. It provides an automated method for identifying and analyzing related sequences (BLAST searches coupled with ClustalW global alignments), and a series of user-configurable rules (the 'Filter') to select sequences; it returns a set of full sequences ready for further analysis.

GraphAlign can analyze the query and sequence set produced by BLAST Filter or any other method. GraphAlign was designed to analyze global pairwise alignments. It can provide information about the similarity between query and subject sequences that is different from that provided by sequence similarity search programs such as BLAST (4) and FASTA (5), which match subject sequences by finding high-quality local alignments. This places different constraints on the alignment process than does local alignment and allows regions of dissimilarity to be explored as well as regions of similarity. GraphAlign employs novel analyses and produces graphical representations of regions of sequence similarity, including subalignments that meet specified thresholds.

BLAST Filter

A BLAST Filter analysis is performed in two steps: (i) sequence similarity analysis and (ii) rule analysis. The rule analysis may be repeated with different rules until the user is satisfied with the sequence set.

Step 1: sequence similarity analysis In this step, a user-supplied query sequence is searched against a selected National Center for Biotechnology Information (NCBI) database by the BLASTP (protein) or BLASTN program. A limit of 1000 sequences is imposed on the BLAST results to reduce

*To whom correspondence should be addressed. Tel: +1 505 646 5031; Fax: +1 505 646 3371; Email: spalding@nmsu.edu computer time and the sizes of output HTML pages. Other BLAST search parameters, however, can be varied. Statistics from the BLAST report are then extracted and stored, and full sequences are retrieved. A ClustalW (6) dynamic programming global pairwise alignment is then performed between the query and every subject sequence matched by BLAST in order to compute the percentages of identities, positives and gaps needed for full-length query–subject rules.

Step 2: rule analysis This step uses the data computed in Step 1 and can be repeated until the user is satisfied with the rule values. There are 15 rules (Table 1) that can be used to 'filter' (i.e. remove) sequences from the full set returned by BLAST. Each rule has a value; more stringent values result in more sequences filtered. When Step 2 is first presented, all rule values are set to defaults that pass all sequences, and a graphical alignment of all high-scoring segment pairs (HSPs) to the query sequence is shown (Figure 1). Unlike the NCBI BLAST report, every HSP for each subject sequence is displayed separately and additional BLAST and global pairwise alignment statistics are shown. For DNA (BLASTN) searches, matches to the reverse complement of the query strand are portrayed differently from forward matches. A mouseover of an HSP bar shows the annotation for that sequence, and clicking on a bar links to that HSP in the BLAST report. The BLAST report is modified to include links between the summary results section at the top and the alignment section, as well as hyperlinks to the GenBank entry for each sequence. A one-line description page is provided that summarizes each matched sequence and also includes hyperlinks to GenBank.

There are two classes of rule that are based on full-length query/subject sequence comparisons (Table 1, rules 1–5): sequence length and global pairwise alignment. If a subject sequence fails any of these rules, it is filtered out. The third class of rule (6–13) is based on BLAST HSP statistics. If at least one HSP passes all these rules, the subject sequence passes this class of rule. The purpose of the final rule class (rules 14,15) is to further reduce the sequence set by removing redundant sequences and limiting the set to a specified size. Redundant sequences are identified by performing an all-against-all ClustalW fast pairwise alignment of subject sequences and removing the lower scoring member of a pair with a ClustalW percentage similarity greater than the user-specified value. This results in a set in which no two members have a percentage similarity above the rule value. Executing this rule with a large number (>200) of sequences may be time-consuming (several minutes or more), so should not be used if immediate feedback is a factor.

Table 1. BLAST Filter rules applied to each query–subject sequence pair

Rule class	Specific rules
Based on sequence length	(1) Minimum sequence length as proportion of query length (2) Maximum sequence length as proportion of query length
Based on ClustalW global pairwise alignment of query with subject	(3) Minimum percentage identity (4) Minimum percentage positives (protein comparisons only) (5) Maximum percentage gaps
Based on BLAST high-scoring segment pair statistics	(6) Minimum Expect (*E*-) value (7) Maximum Expect (*E*-) value (8) Minimum score (bits) (9) Maximum score (bits) (10) Minimum percentage overlap to query (11) Minimum percentage identity (12) Minimum percentage positives (protein comparisons only) (13) Maximum percentage gaps
Final reduction of sequences that pass above rules	(14) Maximum percentage ClustalW similarity (reduce redundancy) (15) Maximum number of sequences (lowest scoring sequences removed)

The power of this step is the ability to change rules, reapply them and immediately view the results. The number of sequences that failed each rule is shown after the rule, and the graphical HSP alignment is color-coded to reflect passed and failed sequences and HSPs, redundant sequences, and those failing the maximum number rule (Figure 1). The filtered sequence set in FASTA format may also be downloaded in this step; the sequences are in the same order as in the BLAST report, i.e. by ascending *E*-values.

Types of application

The simplest use of BLAST Filter is to use the default rules, which returns the full sequences for all BLAST matches. There are many possibilities, however, for customizing the sequence set for a particular application. In the following examples, one could restrict the set to only those sequences with the following properties (numbers refer to rules in Table 1):

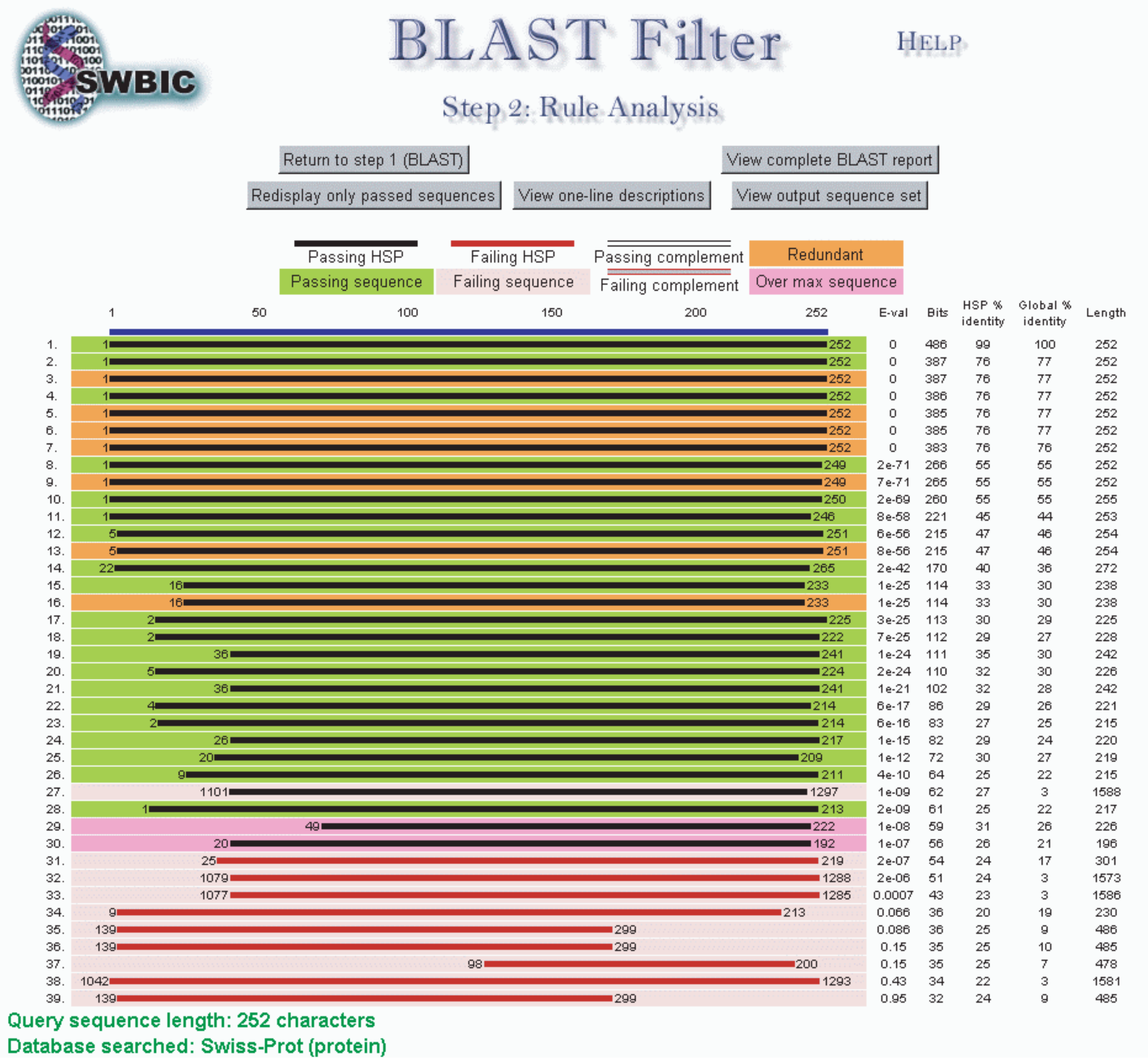
(i) non-redundant, similar, full-length sequences: use rules to control sequence length similarity to the query (1,2), minimum global pairwise percentage identity (3) and maximum percentage similarity within the sequence set (14);

(ii) high-quality HSPs and domain identification: use HSP rules to control maximum *E*-value (7), minimum percentage overlap to the query (10), minimum percentage identity (11) and positives (12), and maximum percentage gaps (13); with a single domain as the query and a high value for rule 10, this would identify sequences containing the complete domain;

(iii) constrained similarity to query: set the rules for either the minimum/maximum *E*-values (6,7) or scores (bits) (8,9) so that sequences are neither too similar nor too dissimilar to the query.

Once a set of rule values has been determined for a particular need, BLAST Filter provides a consistent and objective method of forming a set of related sequences given a single query. This allows for new queries to be used without the need for human 'decision-making', which is not reproducible and cannot be part of automated methods.

Planned enhancements include the following: (i) sequences may be trimmed to their highest scoring HSPs just prior to the final reduction of sequences; (ii) a set of rule values may be downloaded in a file by the user to be uploaded in later sessions so that particular sets of rule values will not have to be re-entered by the user; (iii) addition of PSI-BLAST for protein searches; and (iv) new rules based on text matches to sequence



			E-val	Bits	HSP % identity	Global % identity	Length
1.	1	252	0	486	99	100	252
2.	1	252	0	387	76	77	252
3.	1	252	0	387	76	77	252
4.	1	252	0	386	76	77	252
5.	1	252	0	385	76	77	252
6.	1	252	0	385	76	77	252
7.	1	252	0	383	76	76	252
8.	1	249	2e-71	266	55	55	252
9.	1	249	7e-71	265	55	55	252
10.	1	250	2e-69	260	55	55	255
11.	1	246	8e-58	221	45	44	253
12.	5	251	6e-56	215	47	46	254
13.	5	251	8e-56	215	47	46	254
14.	22	265	2e-42	170	40	36	272
15.	16	233	1e-25	114	33	30	238
16.	16	233	1e-25	114	33	30	238
17.	2	225	3e-25	113	30	29	225
18.	2	222	7e-25	112	29	27	228
19.	36	241	1e-24	111	35	30	242
20.	5	224	2e-24	110	32	30	226
21.	36	241	1e-21	102	32	28	242
22.	4	214	6e-17	86	29	26	221
23.	2	214	6e-16	83	27	25	215
24.	26	217	1e-15	82	29	24	220
25.	20	209	1e-12	72	30	27	219
26.	9	211	4e-10	64	25	22	215
27.	1101	1297	1e-09	62	27	3	1588
28.	1	213	2e-09	61	25	22	217
29.	49	222	1e-08	59	31	26	226
30.	20	192	1e-07	56	26	21	196
31.	25	219	2e-07	54	24	17	301
32.	1079	1288	2e-06	51	24	3	1573
33.	1077	1285	0.0007	43	23	3	1586
34.	9	213	0.066	36	20	19	230
35.	139	299	0.086	36	25	9	486
36.	139	299	0.15	35	25	10	485
37.	98	200	0.15	35	25	7	478
38.	1042	1293	0.43	34	22	3	1581
39.	139	299	0.95	32	24	9	485

Query sequence length: 252 characters

Database searched: Swiss-Prot (protein)

Figure 1. Example of the rule analysis output page of BLAST Filter for a search of a 3-dehydroquinate-dehydratase sequence against the Swiss-Prot database. The option buttons at the top allow the user to return to Step 1 (to change BLAST parameters), declutter the results by displaying data only for passed sequences (this can be undone) and view the BLAST report and output sequence set (FASTA format). Below these is the graphical alignment of HSPs with its legend. Each bar represents the alignment of an HSP with the query and is a link to that HSP in the BLAST report; the bar color indicates whether the HSP passed (black) or failed (red) the HSP statistic rules (Table 1, rules 6–13). The background color for a sequence indicates whether the sequence passed (green) or failed (light red) the first three classes of rule, or failed the redundancy (orange) or maximum number of sequences (purple) rules. The rules, their current values and the number of sequences failing each rule, are included below the alignment section (data not shown). The user may change any rule values and click on the 'Apply changed rules' button to reanalyze and redisplay these results. Also not shown is the page of one-line descriptions, which contains summary information for each sequence and hyperlinks to the GenBank entries.

annotation. Suggestions for additional rules are welcomed, from the research community.

GraphAlign

GraphAlign analyzes global pairwise alignments of nucleotide and protein sequences in novel ways and presents the alignments in a graphical form. The user submits a query sequence and one or more subject sequences, or these sequences may be imported directly from BLAST Filter results. The query sequence is then paired with each subject and a ClustalW 'slow pairwise' (full dynamic programming) gapped alignment with default parameters is performed for each pair. These alignments are then analyzed to find sections (subalignments) of high quality, and graphs of each alignment are prepared that show which parts of the alignments are the most similar.

GraphAlign analyses are novel in two respects. First, regions of similarity and dissimilarity between two sequences are shown by the height of a continuous curve that can be calculated in various ways. Unlike protein domain/motif database servers such as Pfam (7), which show domain matches graphically in the form of colored blocks along sequences, conserved protein domains and motifs in GraphAlign appear as high regions of the curve. Although GraphAlign does not identify specific domains, it can be used in the exploration of similarities among proteins in sets of any type (such as those

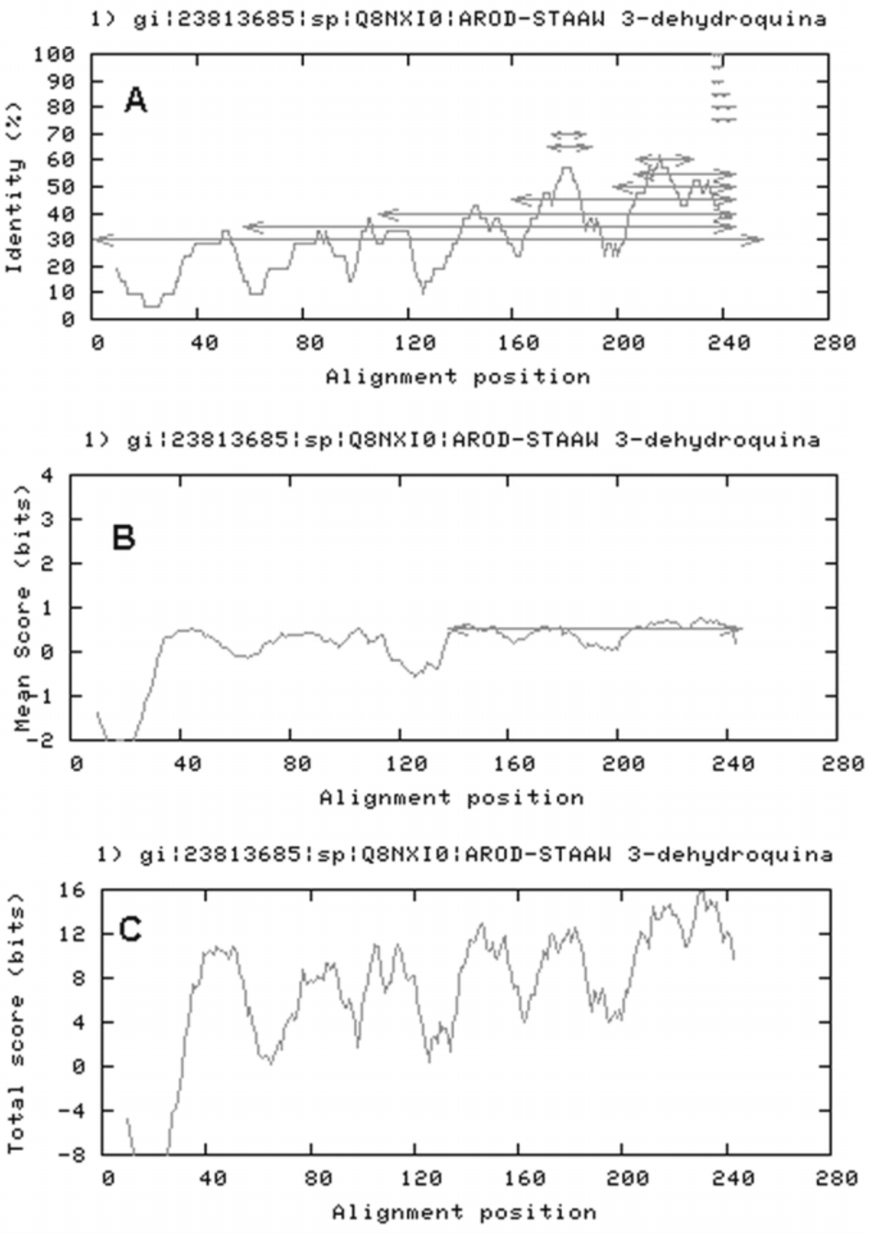


Figure 2. Examples of GraphAlign plots based on an analysis of two 3-dehydroquinate-dehydratase protein sequences. The title of each graph is the definition line of the subject sequence, preceded by the rank of that sequence in the input order of sequences. The value at each x position is calculated from the center of a moving window (here 21 positions) through the alignment. (**A**) Percentage identity curve showing Significant Subalignments (Best Subalignments for percentage identities between 30% and 100% in 5% intervals). Note that regions of highest percentage identity are centered on positions 180 and 240). (**B**) Mean score (bits) curve showing the Critical Subalignment for a Critical Length of 25%; the gap penalty is increased to −4 so that gapped sections of the alignment (centered on positions 19 and 127) become more apparent. (**C**) Total score (bits) curve. The last curve type is a sliding window of total subalignment scores, and therefore the magnitude depends on the window size, unlike the first two types, which normalize the number of identities and sum of scores by the subalignment length.

created by BLAST Filter), including new protein families. GraphAlign can also be used as a general method of exploring similarities among nucleotide sequences. Second, high-quality subalignments of different types are shown graphically, and the subject sequences and their graphs may be sorted by the quality of these subalignments. Like the curves, these subalignments may be calculated in different ways, depending on what the user considers most important in the nature of alignments.

GraphAlign performs its calculations in two stages. The ClustalW alignments are done and results are analyzed and output after the initial submission (stage 1). The user may then click on Modify Parameters (stage 2) to change analysis parameters and redisplay changed output tables and graphs. Stage 2 can be repeated indefinitely.

To understand how GraphAlign works requires explanations of how alignment quality is measured, how the graphical curves are calculated and special types of threshold alignment.

Evaluation of alignment quality

The qualities of alignments (both full and subalignments) are computed in one of three ways:

(i) *Percentage identity*. This measure is calculated by dividing the number of identities (pairs of identical letters) in an alignment by the length of the alignment; gaps are counted as mismatches.

(ii) *Mean score* (protein sequences only). This method assigns a score to each position in the alignment. If the position contains a gap, the gap penalty specified by the user (default −1) is used. If not, the score (in units of bits) for the pair of amino acids is determined from the substitution scoring matrix selected by the user (default BLOSUM30). The mean score for an alignment is the sum of scores divided by the length of the alignment or subalignment.

(iii) *Total score* (protein sequences only). This is the same as mean score, except that is the sum of scores of the alignment. This is the typical measure used by local alignment methods such as FASTA and BLAST, which find local alignments that maximize the sum of scores.

Curves of percentage identity, mean score and total score

Each graph presents (at minimum) a curve that shows the percentage identity, the mean score or total score at each point in the alignment, depending on which evaluation method was selected. Examples of these curves are shown in Figure 2. By default the window is set at 21 alignment positions but may be set to other values to obtain less or more smoothed curves.

Special types of threshold subalignment

GraphAlign also performs calculations to identify regions of the alignment than meet specified thresholds. These regions are determined in novel ways and are defined as follows.

For a given threshold percentage identity, mean score or total score, the *Best Subalignment* is the longest section of the alignment that achieves this value. For example, if the threshold percentage identity is 50%, GraphAlign finds the longest part of the alignment with at least 50% identity. *Significant Subalignments* are Best Subalignments for a series of preset thresholds and may be displayed in the graphs (Figure 2A). Each Significant Subalignment is shown as a horizontal line spanning the region of the alignment and with a y-axis position equal to that of the preset threshold value.

The *Critical Subalignment* is another novel type of threshold alignment and depends on a user-specified value called the *Critical Length*. This length is a percentage of the global alignment length. The *Critical Subalignment* is defined as that region of the alignment with the highest percentage identity or mean or total score that is at least as long as the *Critical Length*. This highest percentage identity or mean or total score is called the *Critical Percentage Identity*, *Critical Mean Score*, or *Critical Total Score*. An example using a Critical Length of 25% with mean score is shown in Figure 2B. This Critical Mean Score can be understood as answering the question 'What is the highest mean score that can be found for a subalignment whose length is equal to or greater than the Critical Length.' Another way to interpret this result is that all subalignments with a mean score >0.5 (the value in this example) are shorter than the Critical Length.

Table 2. GraphAlign inputs

Input option	Description/values
Query sequence	This sequence is paired with each subject sequence
Subject sequences	One or more sequences
Critical Length (%)	A percentage of the alignment length (see text)
Alignment quality evaluation method	Percentage identity, or mean or total score (protein only)
Scoring matrix	A BLOSUM substitution scoring matrix (protein only)
Gap penalty	≤0; protein sequences only, mean or total score method
Subalignments to show	None, Significant Subalignments or Critical Subalignments
Window size for curves	Any odd number up to length of alignment
Sort results by	Many options[a]; table results and graphs are shown in this order
Number of graphs/page	Number of alignment graphs to show in each output page
Number of graph columns/page	Number of columns of graphs to show in each output page
Graph size	Small, medium or large (use with number of columns to fit output on printed pages)

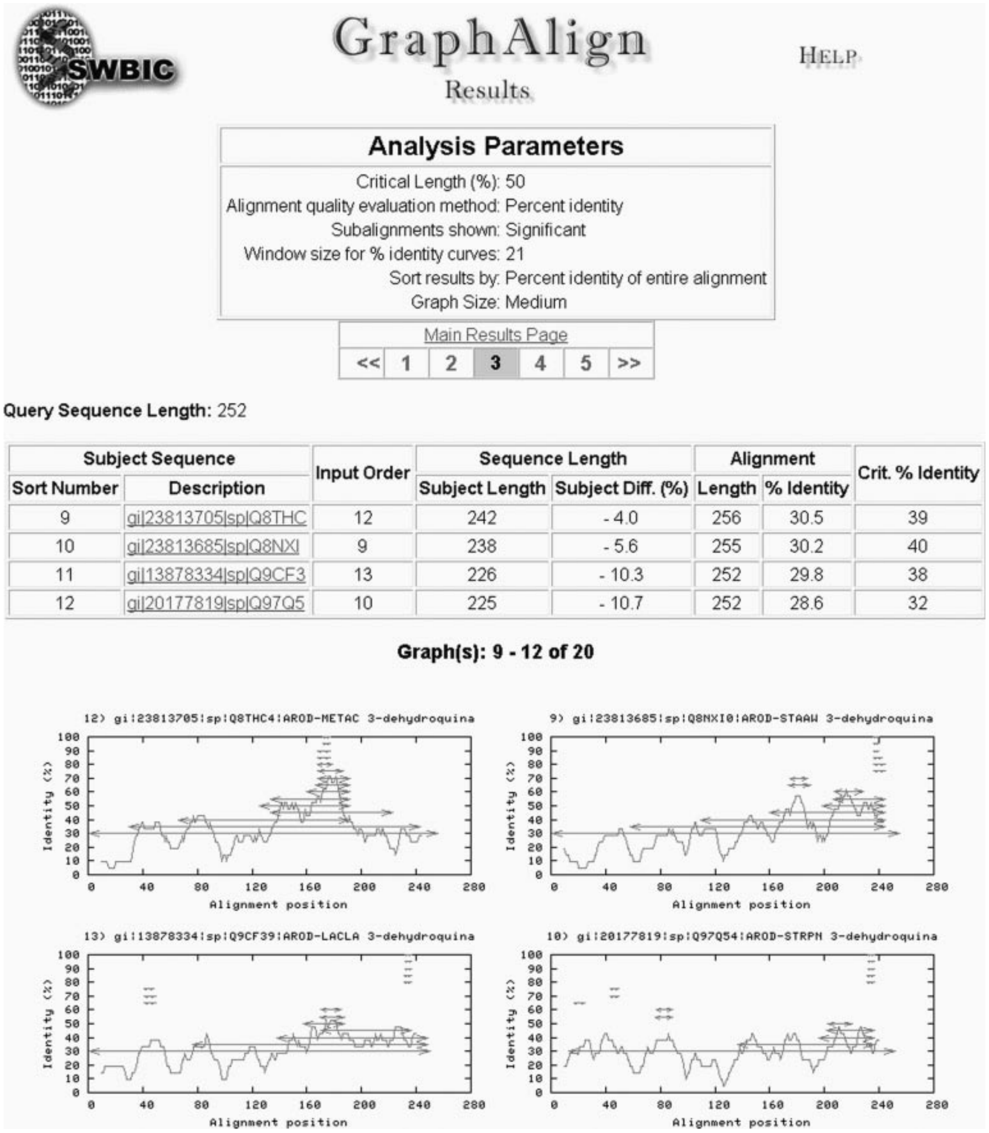
[a] Results may be sorted by these values: global percentage identity or mean or total score, Best Subalignment length, Best Subalignment length/total alignment length (%), Critical Percentage Identity or Mean or Total Score, sequence length difference (%). See text and Help page for further explanations of input options.

Inputs and limits

The inputs to GraphAlign are listed in Table 2. The Best Subalignment percentage identity or mean or total score is also an input, but is prompted for when sorting by Best Subalignment length. Limits on sequences are as follows: 1000 characters per line, 200 subject sequences and no resulting alignment may have more than 10 000 positions.

Output format and interpretation

The results are presented in one or more pages, sorted by the chosen parameter. The number of pages depends on the 'number of graphs per page' value selected. If all of the graphs fit on one page, then all of the results are on one page. If not, the first page consists of a summary table of results for all alignments, and links are provided to navigate through the pages with the graphs. For example, if the user submitted 20 subject sequences and chose to show 4 graphs per page, there would be 6 output pages: the summary table of all results, plus 5 pages of results with tables and graphs. An example of the latter type of output page and its interpretation is given in Figure 3.



GraphAlign

HELP

Results

Analysis Parameters

Critical Length (%): 50
Alignment quality evaluation method: Percent identity
Subalignments shown: Significant
Window size for % identity curves: 21
Sort results by: Percent identity of entire alignment
Graph Size: Medium

Main Results Page

<< 1 2 3 4 5 >>

Query Sequence Length: 252

Subject Sequence		Input Order	Sequence Length		Alignment		Crit. % Identity
Sort Number	Description		Subject Length	Subject Diff. (%)	Length	% Identity	
9	gi|23813705|sp|Q8THC	12	242	- 4.0	256	30.5	39
10	gi|23813685|sp|Q8NXI	9	238	- 5.6	255	30.2	40
11	gi|13878334|sp|Q9CF3	13	226	- 10.3	252	29.8	38
12	gi|20177819|sp|Q97Q5	10	225	- 10.7	252	28.6	32

Graph(s): 9 - 12 of 20

12) gi!23813705!sp!Q8THC4!AROD-METAC 3-dehydroquina
9) gi!23813685!sp!Q8NXI0!AROD-STAAW 3-dehydroquina
13) gi!13878334!sp!Q9CF39!AROD-LACLA 3-dehydroquina
10) gi!20177819!sp!Q97Q54!AROD-STRPN 3-dehydroquina
Identity (%)
0
10
20
30
40
50
60
70
80
90
100
0
40
80
120
160
200
240
280
Alignment position

Figure 3. Example of a GraphAlign result page from an analysis of the 3-dehydroquinate-dehydratase sequence set produced by BLAST Filter (Figure 1). The top of each output page shows the current analysis parameter options followed by a summary table. The columns of the summary tables include a variety of statistics that are described fully in the Help page. The alignment graphs are shown below the table; the interpretation of graphs is discussed in Figure 2. This example illustrates four sequences with ~30% identity to the query. PROSITE (8) and InterPro (9) searches of the query revealed one significant motif/pattern (DEHYDROQUINASE_1, the active site), which starts at about position 115 in these alignments. This motif is 32–34 residues long with a variable section of 18–20 residues. Because of the variability of this motif, it is not apparent in the graphs. Knowledge of this motif, however, would not have revealed the regions of high similarity to the query in these subject sequences, as indicated by the horizontal significant subalignment lines. Not only are there regions with much higher similarity than the active site, but there is variability among these sequences in the positions of highest similarities. To examine a particular region, the user clicks on a link in the 'Description' column, which shows the ClustalW pairwise alignment for that subject sequence with alignment positions numbered (data not shown).

ACKNOWLEDGEMENTS

Thanks to Ciju John and Chris Hoxworth for the server implementations. R. L. Bernstein provided testing and advice. This work was funded as part of the Army Research Organization grant #DAAH04-96-1-0415 for the Southwest Biotechnology and Informatics Center, and NIH grant S06 GM08136 to P.L. and J.S.

REFERENCES

1. Phillips,A., Janies,D. and Wheeler,W. (2000) Multiple sequence alignment in phylogenetic analysis. *Mol. Phylogenet. Evol.*, **16**, 317–330.
2. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
3. Cuff,J.A. and Barton,G.J. (1999) Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*, **34**, 508–519.
4. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
5. Pearson,W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.*, **183**, 63–98.
6. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

7. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
8. Sigrist,C.J., Cerutti,L., Hulo,N., Gattiker,A., Falquet,L., Pagni,M., Bairoch,A. and Bucher,P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform.*, **3**, 265–274.
9. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.